Objetivo

Utilizar os dados disponíveis e a linguagem R para tentar encontrar fatores que tenham correlação e/ou influência quando meninas decidem qual curso superior querem fazer e se vão ou não fazer um curso na área de computação.

Introdução

Durante os anos de 2011 até 2014, a Professora Marístela de Holando aplicou um formulário buscando encontrar fatores que são impactantes quando um menina quer escolher um curso de computação. Isso foi motivado pelo fato de que o número de meninas que ingressão nos cursos dessa área é muito menor do que a quantidade de meninos que buscam cursar os mesmos cursos.

O professor Vinícius Borges disponibilizou uma planilha de dados de uma pesquisa realizada de 2011 a 2014 na Semana Nacional de Ciência e Tecnologia (SNCT) em Brasília e tal pesquisa buscava descobrir um pouco mais sobre as meninas do ensino médio de Brasília e por que o baixo nível de interesse nas áreas de Computação.

O trabalho consistirá em transformar e carregar os dados (o processo de extração foi previamente feito). A partir das informações, espera-se obter um insight sobre a atual situação das mulheres na computação.

Pacotes usados

```
require(dplyr)
require(party)
```

Limpeza dos Dados

Inicialmente, o Data Frame possui as seguintes informações armazenadas.

```
workingdata <- read.csv("data.csv", header = TRUE, stringsAsFactors=FALSE)
names(workingdata)
```

```
##  [1] "Year"
##  [2] "Gender"
##  [3] "Educational.Stage"
```

```
##  [4] "Field.Of.Interest"
##  [5] "Would.Enroll.In.CS"
##  [6] "Q1"
##  [7] "Q2"
##  [8] "CS.Only.Teaches.To.Use.Software"
##  [9] "CS.Uses.Little.Math"
## [10] "Most.CS.Students.Are.Male"
## [11] "CS.Requires.Knowledge.In.Computers"
## [12] "Higher.Education.Required.To.Work.In.CS"
## [13] "Family.Approves.CS.Major"
## [14] "CS.Has.Low.Employability"
## [15] "CS.Work.Has.Long.Hours"
## [16] "CS.Fosters.Creativity"
## [17] "CS.Is.Prestigious"
## [18] "CS.Provides.Good.Wages"
## [19] "CS.Enables.Interdisciplinary.Experiences"
## [20] "Uses.Computer.At.Home"
## [21] "Uses.Computer.At.Relatives.House"
## [22] "Uses.Computer.At.Friends.House"
## [23] "Uses.Computer.At.School"
## [24] "Uses.Computer.At.Work"
## [25] "Uses.Computer.At.Lan.House"
## [26] "Uses.Computer.At.Library"
## [27] "Uses.Computer.At.Digital.Inclusion.Center"
## [28] "Has.Used.Text.Editor"
## [29] "Has.Used.Image.Editor"
## [30] "Has.Used.Spreadsheet"
## [31] "Has.Used.Database"
## [32] "Has.Used.Internet"
## [33] "Has.Used.Social.Network"
## [34] "Has.Used.Email"
## [35] "Has.Used.Games"
## [36] "Has.Used.For.Creating.Web.Pages"
## [37] "Has.Used.For.Development"
## [38] "Has.Used.Other.Softwares"
```

Primeiro passo, sabe-se que os alunos que responderam os questinários colocaram seu “Gênero”.

```
workingdata %>%
  group_by(Gender) %>%
  summarize(total = n())
```

```
## # A tibble: 3 × 2
##   Gender total
##    <chr> <int>
## 1           14
## 2      F  3680
## 3      M    13
```

Como é possível notar, há a presença de 13 meninos e 14 pessoas não informaram o sexo, portanto, essas 27 pessoas devem ser removidas da análise.

```
workingdata <- workingdata[workingdata$Gender == 'F',]
```

Atualmente os dados encontram-se em sua maioria em formato de string, tendo como respostas "Yes", "No" e "Maybe". Para que se torne mais fácil de manipular, iremos transformar tais strings em zeros ("No"), uns ("Yes") e dois ("Maybe").

```
workingdata[,5:38][workingdata[,5:38] == "No"] = 0
workingdata[,5:38][workingdata[,5:38] == "Yes"] = 1
workingdata[,5:38][workingdata[,5:38] == "Maybe"] = 2
```

Os campos de interesse constituem basicamente de "Human Sciences", "Biology-Health Sciences" e "Exact Sciences", nomes relativamente grandes, portanto abreviaremos para 0, 1 e 2 respectivamente.

```
workingdata$Field.Of.Interest[workingdata$Field.Of.Interest == "Human Sciences"]
workingdata$Field.Of.Interest[workingdata$Field.Of.Interest == "Biology-Health S
workingdata$Field.Of.Interest[workingdata$Field.Of.Interest == "Exact Sciences"]
```

Uma vez transformadas as informações, elas serão convertidas em fatores para melhor manipulá-las.

```
cols <- c(4,5,8:38)
workingdata[cols] <- lapply(workingdata[cols], factor)
```

Análise exploratória

Agora que o dado encontra-se bem formatado, é possível explorá-lo com maior facilidade.

** Verificar porcentagens

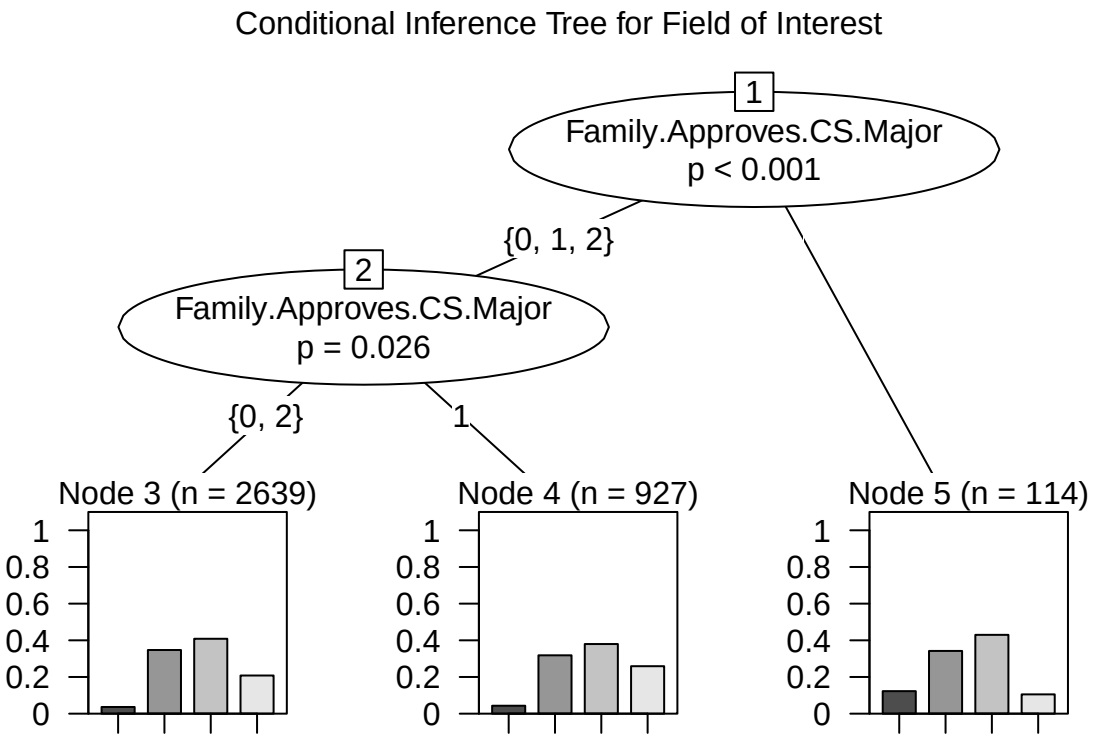
** Verificar possíveis associações para a parte seguinte

**

Predição

Para o próximo passo, tentaremos prever a área de interesse usando as respostas dadas às ultimas perguntas, além de verificar qual pergunta é mais pertinente à previsão da área de interesse.

```
tree <- ctree(Field.Of.Interest ~ Family.Approves.CS.Major + CS.Is.Prestigious,
   data=workingdata)
plot(tree, main="Conditional Inference Tree for Field of Interest")
```



Conclusão

** Comentar sobre os dados da análise ** Comentar sobre a predição ** Comentar sobre a impossibilidade de saber se as estudantes foram ou não para CIC